# Applying Deep Learning Techniques to Extract Ultrasonic Mouse Vocalizations

Drew Kristensen ‘18, Dr. Adam A. Smith (Advisor)
Department of Mathematics and Computer Science, University of Puget Sound

## What is it?

### Ultrasonic Vocalizations:

Mice are useful in developing treatments for various diseases and disorders, and while it is relatively easy to understand how they physically respond to a treatment, understanding how it impacts their affective state proves to be a difficult task. To do this, scientists use the mice’s vocalizations to attempt to gain insight into how they are responding. These vocalizations are typically emitted between 30 and 120 kHz, qualifying the calls as ultrasonic as they are above the limit of human hearing of 20 kHz.

### Artificial Neural Networks:

Artificial neural networks (ANNs) are comprised of ‘layers’ of artificial ‘neurons’ which are connected to the layers adjacent to itself. Each of these ‘neurons’ take in a weighted sum of some inputs, pass the result through some activation function, and return the final value as its output. The weights for each neuron can be changed through a process known as training, in which the network is shown example inputs and what the output should be. By propagating the difference of the predicted output and the ground truth through the network, the weights are altered in a way to better achieve the result.

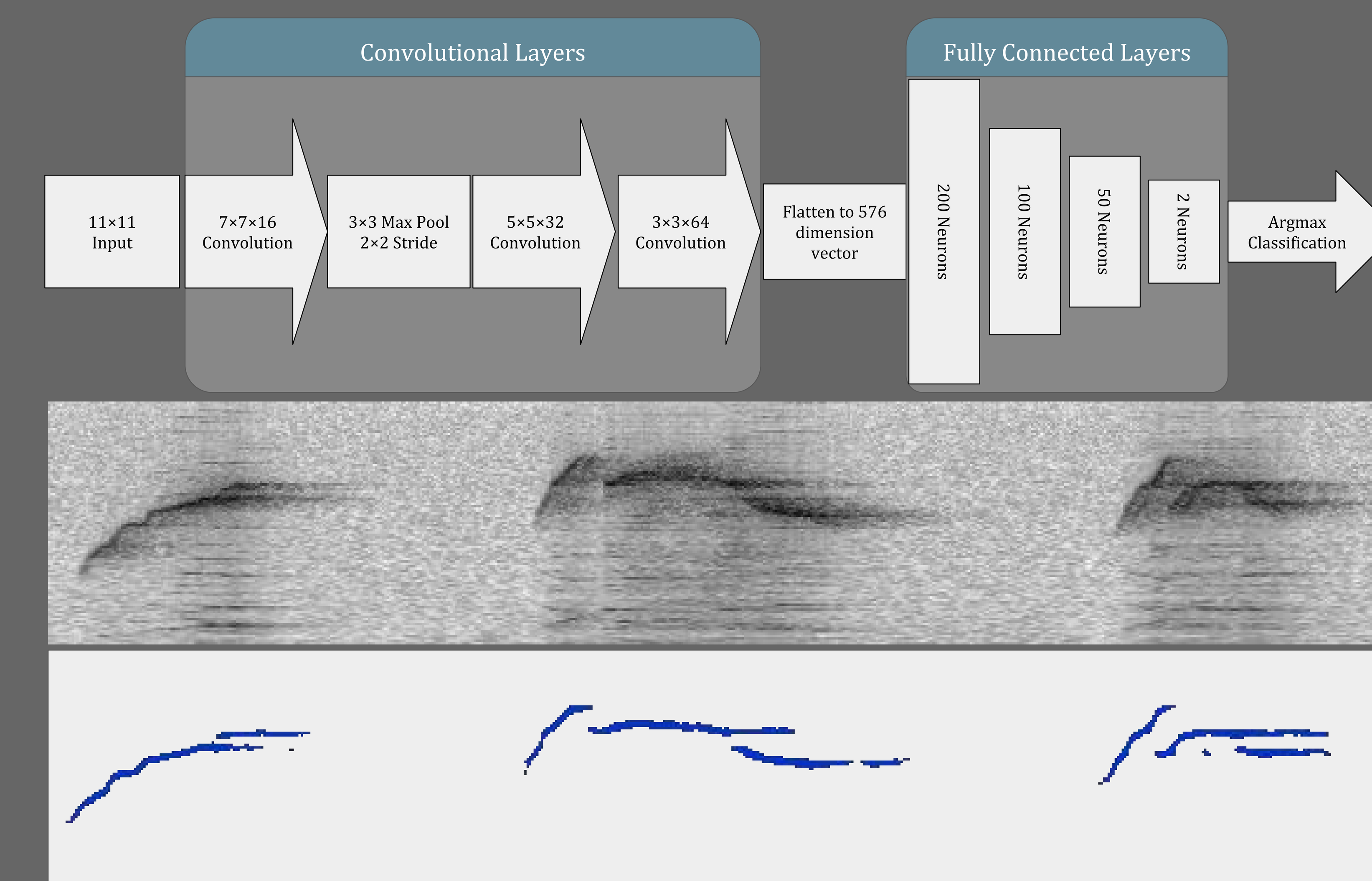
Convolutional neural networks (CNNs) are ANNs where the initial layers pass weighted feature maps across an image input and return the results of the weighted features as the input to the next layer.

### Deep Learning:

Deep Learning is an idea in Machine Learning in which artificial neural networks are used with more than one hidden layer, making it a ‘deep’ network. Deep neural networks have been shown to perform better than their ‘shallower’ counterparts across many fields such as Natural Language Processing [1], Computer Vision [2], and Generative Models [3].

## Results

The final architecture that we selected consisted of three convolutional layers followed by linear rectifier activation functions, with one max pooling layer between the first and second convolutions. On top of the convolutions, we had three fully connected layers of sizes 200, 100, and 50 respectively, all using the linear rectifier activation function.
To measure the success of our network, we computed the $F_1$ scores - a combination of precision and recall - across multiple recordings. We computed the pixel-wise, time-wise, and call-wise F1 scores for the CNN, a normal feed forward neural network, and the Song filtering algorithm.

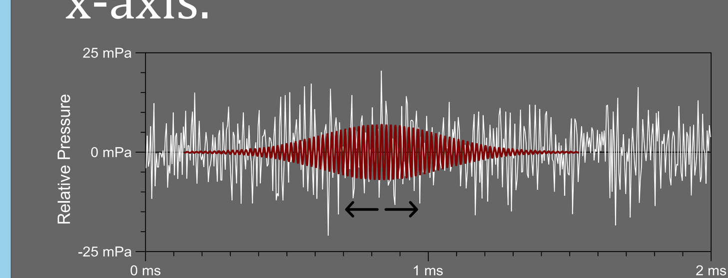




Top: Convolutional Network Architecture
Middle: Example section of a scaleogram
Bottom: corresponding CNN output from section

It is clear that the CNN architecture performed better than the other two extractions methods that were used as controls.
We hope that this new method for extraction can be applied to real world examples in research labs as it could save time for researchers as well as improve accuracy of their results.

## How can we visualize audio?

In order to visualize the audio files, we use a process known as continuous wavelet transformations (CWT), in which we pass different scales of a mother wavelet function across the raw audio, multiplying the intensity of the sound waves with the corresponding crests and troughs of the scaled wavelet. In this way, we can generate an accurate visualization of the recording known as a scaleogram, with frequency scaled logarithmically on the y-axis and time on the x-axis.




Example of a Continuous Wavelet Transformation

Example scaleogram output from the CWT

## Works Cited

[1] Cho et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation arXiv:1406.1078v3
[2] Szegedy et al. Going deeper with convolutions arXiv:1409.4842
[3] Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks arXiv:1511.06434

## Acknowledgments

I would like to thank the McCormick grant foundation for providing me the funding to be able to pursue this project, and to MouseTube for providing the vocalization recordings. Lastly, I would also like to thank my research advisor, Dr. Adam A. Smith for his invaluable help and insight on this project.